Practice Problems 2 (with solutions) - Chapter 4 - CSC/DSC 262/462

1. Two dice are tossed independently. Let X_1, X_2 be random variables representing the two outcomes, each from sample space $S_X = \{1, 2, 3, 4, 5, 6\}$. Derive the proability mass function of the following random variables:

   (a) $X = X_1 + X_2$,

   (b) $X = \max(X_1, X_2)$.

SOLUTION We have a random experiment with 36 equiprobable outcomes from sample space

$$S = \{(i, j) : i = 1, \ldots, 6, \;\; j = 1, \ldots, 6\}.$$

The random variables X_1, X_2 are determined by representing the outcome as (X_1, X_2).

(a) The PMF for X is given by

$$p_i = P(X = i) = P(X_1 + X_2 = i).$$

The support of X is $\mathcal{S}_X = \{2, 3, \ldots, 11, 12\}$. Then the PMF is given by

$$\begin{aligned}
p_2 = P(X = 2) &= P\left((X_1, X_2) \in \{(1,1)\}\right) = 1/36, \\
p_3 = P(X = 3) &= P\left((X_1, X_2) \in \{(1,2), (2,1)\}\right) = 2/36, \\
p_4 = P(X = 4) &= P\left((X_1, X_2) \in \{(1,3), (2,2), (3,1)\}\right) = 3/36, \\
p_5 = P(X = 5) &= P\left((X_1, X_2) \in \{(1,4), (2,3), (3,2), (4,1)\}\right) = 4/36, \\
p_6 = P(X = 6) &= P\left((X_1, X_2) \in \{(1,5), (2,4), (3,3), (4,2), (5,1)\}\right) = 5/36, \\
p_7 = P(X = 7) &= P\left((X_1, X_2) \in \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}\right) = 6/36, \\
p_8 = P(X = 8) &= P\left((X_1, X_2) \in \{(2,6), (3,5), (4,4), (5,3), (6,2)\}\right) = 5/36, \\
p_9 = P(X = 9) &= P\left((X_1, X_2) \in \{(3,6), (4,5), (5,4), (6,3)\}\right) = 4/36, \\
p_{10} = P(X = 10) &= P\left((X_1, X_2) \in \{(4,6), (5,5), (6,4)\}\right) = 3/36, \\
p_{11} = P(X = 11) &= P\left((X_1, X_2) \in \{(5,6), (6,5)\}\right) = 2/36, \\
p_{12} = P(X = 12) &= P\left((X_1, X_2) \in \{(6,6)\}\right) = 1/36.
\end{aligned}$$

(b) The PMF for X is given by

$$p_i = P(X = i) = P(\max(X_1, X_2) = i).$$

The support of X is $\mathcal{S}_X = \{1, 2, 3, 4, 5, 6\}$. Then the PMF is given by

$$\begin{aligned}
p_1 = P(X = 1) &= P\left((X_1, X_2) \in \{(1,1)\}\right) = 1/36, \\
p_2 = P(X = 2) &= P\left((X_1, X_2) \in \{(1,2), (2,1), (2,2)\}\right) = 3/36, \\
p_3 = P(X = 3) &= P\left((X_1, X_2) \in \{(1,3), (2,3), (3,3), (3,1), (3,2)\}\right) = 5/36, \\
p_4 = P(X = 4) &= P\left((X_1, X_2) \in \{(1,4), (2,4), (3,4), (4,4), (4,3), (4,2), (4,1)\}\right) = 7/36,
\end{aligned}$$

$$\begin{aligned} p_5 = P(X=5) &= P((X_1,X_2) \in \{(1,5),(2,5),(3,5),(4,5),(5,5),\ldots \\ &\quad \ldots,(5,4),(5,3),(5,2),(5,1)\}) = 9/36, \\ p_6 = P(X=6) &= P((X_1,X_2) \in \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6),\ldots \\ &\quad \ldots,(6,5),(6,4),(6,3),(6,2),(6,1)\}) = 11/36. \end{aligned}$$

2. A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} c(x+1) & ; \quad x \in [1,3] \\ 0 & ; \quad otherwise \end{cases}.$$

(a) Determine c.

(b) Determine $P(X \le 2)$.

SOLUTION

(a) We have

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_1^3 c(u+1)du = c(u^2/2+u)\Big|_1^3 = c \times 6.$$

This means $c = 1/6$.

(b) We have

$$P(X \le 2) = \int_{-\infty}^{2} f_x(u)du = \int_1^2 (u+1)/6\, du = (u^2/2+u)/6\,\Big|_1^2 = 5/12.$$

3. A bin contains m white and n black balls. A random selection of $k \le m+n$ balls is made (this is referred to as *sampling without replacement*). Let X be the number of white balls among the k selected. This is known as a *hypergeometric random variable*, which we denote $X \sim hyper(m,n,k)$.

(a) Using principles of combinatorics, derive a general expression for the PMF of X. Make sure to state exactly the support $\mathcal{S}_X$ of X.

(b) Define a sequence of Bernoulli random variables $U_1,\ldots,U_k$, setting $U_i = 1$ if the ith selected ball is white. Expressing X as their sum, determine the mean and variance of X.

(c) Suppose we make a selection of k balls in the same manner, except that the balls are replaced immediately after being selected, and may be selected again (this is referred to as *sampling with replacement*). Let Y be the total number of white balls selected. What distribution does Y have? Show that $E[X] = E[Y]$ and determine the ratio $var[X]/var[Y]$. Verify that $var[X] \le var[Y]$ for $k \ge 1$ and $var[X] < var[Y]$ for $k > 1$.

(d) Based on the comparisons of part (c), under what conditions can the distribution of Y be used to approximate the distribution of X?

(e) A lake contains N fish. Suppose J fish are caught, tagged, then released. After a period of time, K fish are caught (assume these K fish are distinct). Suppose X of these have been previously tagged. Assuming both samples are random samples, we have

$$X \sim hyper(J, N - J, K).$$

Derive an expression using X, J, K, denoted $\hat{N}$, which can used as an estimate of total population size N. Use $E[X]$ as a guide. This method is known as *mark and recapture*, and is commonly used to estimate population sizes.

(f) Since X is random, we would like to know how close $\hat{N}$ is to N. Once way to do this is to use a *confidence set* CS of *confidence level* $1 - \alpha$. Set x_{obs} to be the observed value of X. Then let N^* be a possible value of N. Then

$$N^* \in CS \quad \text{if and only if} \quad P(Y \leq x_{obs}) > \alpha/2 \;\text{ and }\; P(Y \geq x_{obs}) > \alpha/2, \qquad (1)$$

where

$$Y \sim hyper(J, N^* - J, K).$$

It may be shown that CS will consist of all integers between some lower and upper bounds, that is,

$$CS = \{N : N_L \leq N \leq N_U\}$$

for some N_L, N_U. Then

$$P(N \in CS) \geq 1 - \alpha,$$

so that the confidence set contains the true value of N with a probability of at least $1 - \alpha$. Write an `R` function which accepts (X, J, K, α) as input, and outputs the bounds N_L, N_U for a confidence set CS for N of confidence level $1 - \alpha$. To do this, set N_U, N_L to be the maximum and minimum values of N^*, respectively, that satisfy condition (1). If you use a search algorithm, a good starting point would be $\hat{N}$ (rounded off to the nearest integer), which would be within the bounds N_L, N_U. Make use of `R` function `phyper()`.

(g) Use your function to determine a confidence set for N when $X = 13$, $J = 200$, $K = 100$, $\alpha = 0.05$.

SOLUTION

(a) We may temporarily label the balls, so that they are all distinct. Then the total number of selections is

$$D = \binom{m+n}{k}.$$

To enumerate the selections for which $X = i$ use the *rule of product.*

1. Selection combination of i from m white balls, $n_1 = \binom{m}{i}$.

2. Selection combination of $k-i$ from n black balls, $n_2 = \binom{n}{k-i}$.

There are

$$N = n_1 \times n_2 = \binom{m}{i}\binom{n}{k-i}$$

such combinations. We then have the general expression for the PMF:

$$p_X(i) = P(X = i) = \frac{N}{D} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}.$$

To derive the support $\mathcal{S}_X$ we note that the number of white and black balls selected (i and $k-i$, respectively) must satisfy the following inequalities:

$$\begin{aligned} 0 &\leq i \leq m \\ 0 &\leq k-i \leq n \text{ or } k-n \leq i \leq k, \end{aligned}$$

so that the support is given by

$$\mathcal{S}_X = \{i : \max(0, k-n) \leq i \leq \min(k, m)\}.$$

(b) There are m white balls from a total of $m+n$. Therefore, $P(U_i = 1) = p = m/(m+n)$, and $E[U_i] = p$ for $i = 1, \ldots, k$. Therefore

$$E[X] = \sum_{i=1}^{k} E[U_i] = kp = \frac{km}{m+n}.$$

Since $U_i \sim bern(p)$, we have variance

$$var[U_i] = \sigma_i^2 = p(1-p) = \frac{m}{m+n}\left[1 - \frac{m}{m+n}\right] = \frac{mn}{(m+n)^2}.$$

However, the random variables U_i are not independent. Note that for any $i \neq j$ the product U_iU_j is also a Bernoulli random variable, with $U_iU_j = 1$ if and only if the ith and jth ball are both white. If we condition on the event $U_j = 1$, we effectively remove a white ball before making the next selection. Therefore,

$$P(U_i = 1 \mid U_j = 1) = \frac{m-1}{m+n-1}$$

so that

$$E[U_iU_j] = P(U_i = 1, U_j = 1) = P(U_i = 1 \mid U_j = 1)P(U_j = 1) = \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right)$$

and

$$\begin{aligned} cov[U_iU_j] &= E[U_iU_j] - E[U_i]E[U_j] \\ &= \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right) - \left(\frac{m}{m+n}\right)^2 \\ &= \frac{m}{m+n}\left[\frac{m-1}{m+n-1} - \frac{m}{m+n}\right] \\ &= -\frac{mn}{(m+n)^2(m+n-1)}. \end{aligned}$$

We use the expression

$$\begin{aligned} var[X] &= \sum_i \sigma_i^2 + 2\sum_{i<j}\sigma_{ij} \\ &= \frac{kmn}{(m+n)^2} - \frac{k(k-1)mn}{(m+n)^2(m+n-1)} \\ &= \frac{kmn}{(m+n)^2}\left(\frac{m+n-k}{m+n-1}\right). \end{aligned}$$

(c) Selections are now independent, with constant probability $p = m/(m+n)$ of selecting a white ball for each draw. Therefore $Y \sim bin(k, m/(m+n))$. Then

$$E[Y] = \frac{km}{m+n} = E[X],$$

and

$$var[Y] = kp(1-p) = \frac{kmn}{(m+n)^2}.$$

This means

$$\frac{var[X]}{var[Y]} = \frac{m+n-k}{m+n-1},$$

and the inequalities follow directly.

(d) The means of X and Y are equal for all parameters m, n, k. Otherwise, based on the ratio $var[X]/var[Y]$ the variances are approximately equal if k is small compared to $m+n$. In this case, the binomial distribution may be used to approximate the hypergeometric distribution.

(e) We have J 'white balls', $N-J$ 'black balls', and a selection without replacement of size K. This means

$$X \sim hyper(J, N-J, K).$$

We have

$$E[X] = \frac{KJ}{N},$$

and so

$$N \approx \frac{KJ}{X}.$$

(f) The condition (1) can be expressed in terms of the CDF F_Y:

$$P(Y \leq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs}) > \alpha/2$$

and

$$P(Y \geq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs}-1) < 1-\alpha/2.$$

Then the following function produces the confidence set:

```
cs.hyper = function(x,j,k,alpha) {

# To find NL:
# start at rounded estimate, decrement n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x,j,n-j,k) > alpha/2) {n = n-1}
NL = n+1

# To find NU:
# start at rounded estimate, increment n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x-1,j,n-j,k) < (1-alpha/2)) {n = n+1}
NU = n-1

cs = c(NL=NL,NU=NU)
return(cs)
}
```

(g) We have output:

```
> j = 200
> k = 100
> alpha = 0.05
> x = 13
> cs.hyper(x,j,k,alpha)
  NL   NU
 963 2778
```

4. A random variable X possesses the following density function for some constant c:

$$f_X(x) = \begin{cases} cx^3 & ; \quad x \in [0,4] \\ 0 & ; \quad otherwise \end{cases} .$$

(a) Determine c.

(b) Determine CDF $F_X(x)$.

SOLUTION

(a) We have

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_0^4 cu^3 du = \left.\frac{cu^4}{4}\right|_0^4 = c \times 64.$$

This means $c = 1/64$.

(b) For $x \leq 0$ we have $F_X(x) = 0$, and for $x \geq 4$ we have $F_X(x) = 1$. For $x \in [0,4]$ we have

$$F_X(x) = \int_{-\infty}^{x} u^3/64du = \frac{x^4}{256},$$

so that

$$F_X(x) = \begin{cases} 0 & ; \quad x \leq 0 \\ \frac{x^4}{256} & ; \quad x \in (0,4] \\ 1 & ; \quad x > 4 \end{cases}.$$

5. Suppose X is a nonnegative random variable, that is, $P(X \geq 0) = 1$. Assume that X is a discrete random variable with sample space $S_X = \{0,1,2,\ldots\}$. Show that

$$E[X] = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

Hint: Express X as a sum of Bernoulli random variable.

SOLUTION

(a) Set $U_n = 1$ if $n \leq X$ and $U_n = 0$ otherwise. Then

$$X = \sum_{i=1}^{\infty} U_i,$$

and so

$$E[X] = \sum_{n=1}^{\infty} E[U_n] = \sum_{n=1}^{\infty} P(X \geq n) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

6. Suppose $X_1, \ldots, X_n$ are independent random variables with a common CDF F_X.

(a) Show that the CDF of $Y = \max(X_1, \ldots, X_n)$ is given by $F_Y(t) = F_X^n(t)$.

(b) Suppose $X_1 \sim unif(0,1)$. Derive the density function and mean of Y.

SOLUTION

(a) We have, by independence,

$$F_Y(t) = P(Y \leq t) = P\left(\cap_i \{X_i \leq t\}\right) = \prod_i P(\{X_i \leq t\}) = F_X^n(t).$$

(b) For the $unif(0,1)$ distribution we have

$$F_X(t) = \begin{cases} 0 & ; \quad t < 0 \\ t & ; \quad t \in [0,1) \\ 1 & ; \quad t \geq 1 \end{cases}$$

so

$$F_Y(t) = \begin{cases} 0 & ; \quad t < 0 \\ t^n & ; \quad t \in [0,1) \\ 1 & ; \quad t \geq 1 \end{cases} .$$

The density function is the derivative of the CDF, so (apart from $t = 0$ and $t = 1$)

$$f_Y(t) = \frac{dF_Y(t)}{dt} = \begin{cases} nt^{n-1} & ; \quad t \in [0,1] \\ 0 & ; \quad ow \end{cases} .$$

and we have

$$E[Y] = \int_{t=0}^{1} tnt^{n-1}dt = \int_{t=0}^{1} nt^n dt = \frac{n}{n+1}t^{n+1}\Big|_0^1 = \frac{n}{n+1}.$$

7. A circle has radius R, circumference C and area A.

   (a) If $R \sim exp(1)$ derive the density function for C and A. Which of these has an exponential distribution?

   (b) If $R \sim unif(0,1)$ derive the density function for C and A. Which of these has a uniform distribution?

SOLUTION Either the method of Section 4.5.1 (CDF method) or Section 4.5.2 (One-to-one method) can be used. In fact, as shown in Section 4.5.2, they will be essentially the same method when increasing transformations are used. Below, we use the method of Section 4.5.2. We have $C = 2\pi R$ and $A = \pi R^2$. Then use transformation rule

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y)),$$

for transformation $Y = g(X)$. We have transformations

$$\begin{aligned} C &= g_C(R) = 2\pi R, \\ g_C^{-1}(c) &= c/(2\pi), \\ \frac{dg_C^{-1}(c)}{dc} &= (2\pi)^{-1}, \text{ and if} \\ A &= g_A(R) = \pi R^2, \text{ we have} \\ g_A^{-1}(a) &= \sqrt{a/\pi}, \\ \frac{dg_A^{-1}(a)}{dr} &= 2^{-1}(\pi a)^{-1/2}. \end{aligned}$$

(a) We have support $[0, \infty)$ and density function $f_R(r) = \exp(-r)$ for R. The support of C and A is also $[0, \infty)$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1} \exp\left(-c/(2\pi)\right), \;\; c \geq 0, \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2} \exp\left(-\sqrt{a/\pi}\right), \;\; a \geq 0, \end{aligned}$$

with $f_C(c) = 0$ and $f_A(a) = 0$ for $c, a < 0$. Note that $C \sim exp((2\pi)^{-1})$.

(b) We have support $[0, 1]$ and density function $f_R(r) = I\{r \in [0, 1]\}$ for R. The support of C is now $[0, 2\pi]$ and the support of A is now $[0, \pi]$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1} I\{c/(2\pi) \in [0, 1]\} = (2\pi)^{-1} I\{c \in [0, 2\pi]\} \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2} I\{a \in [0, \pi]\}. \end{aligned}$$

Note that $C \sim unif[0, 2\pi]$. These densities are defined on the entire real line, and the support is implicit in the indicator functions. We can also write:

$$f_C(c) = \begin{cases} (2\pi)^{-1} & ; \;\; c \in [0, 2\pi] \\ 0 & ; \;\; ow \end{cases}$$

and

$$f_A(a) = \begin{cases} 2^{-1}(\pi a)^{-1/2} & ; \;\; a \in [0, \pi] \\ 0 & ; \;\; ow \end{cases} .$$

8. Assume that over a 25 year period the mean height of adult males increased from 175.5 cm to 179.1 cm, with the standard deviation remaining constant at $\sigma = 5.84$. Suppose the minimum height requirement to join the police force remained unchanged at 172 cm. Assume the heights are normally distributed. You can use `R` to do your calculations.

(a) What proportion of adult males would not meet the minimum height requirement at the beginning and end of the 25 year period?

(b) To what value should the minimum height be changed after 25 years in order to maintain the same proportion which meet the height requirement?

(c) What proportion of adult males would not have met this updated requirement 25 years ago?

(d) Repeat the first three questions using the same values, except that we'll assume that the standard deviation has increased from 5.84 to 10.0 over the 25 year period.

SOLUTION Set $\mu_1 = 175.5$, $\mu_2 = 179.1$, $\sigma = 5.84$, $\sigma_{new} = 10.0$. Then let $X_1 \sim N(\mu_1, \sigma^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, $X_3 \sim N(\mu_3, \sigma^2_{new})$. Also set $Z \sim N(0, 1)$.

(a) We have

$$p_1 = P(X_1 \leq 172) = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{172 - \mu_1}{\sigma}\right) = P(Z \leq -0.5993) \approx 0.274,$$

$$p_2 = P(X_2 \le 172) = P\left(\frac{X_2 - \mu_2}{\sigma} \le \frac{172 - \mu_2}{\sigma}\right) = P(Z \le -1.2158) \approx 0.112,$$

so that the respective proportions are p_1 and p_2.

(b) Let q_1 be the minimum height required after 25 years. We need the p_1 quantile. For a standard normal distribution this is

$$p_1 = P(Z \le Z_{p_1})$$

which is solved by $Z_{p_1} = -0.5993$. Then the p_1 quantile for the $N(\mu_2, \sigma^2)$ distribution is

$$q_1 = X_{p_1} = \mu_2 + Z_{p_1}\sigma \approx 179.1 + (-0.5993) \times 5.84 = 175.6.$$

The new minimum height should be $q_1 = 175.6$.

(c) Let p_3 be the proportion not meeting the new minimum requirement q_1 25 years ago. Then

$$p_3 = P(X_1 \le q_1) = P\left(\frac{X_1 - \mu_1}{\sigma} \le \frac{175.6 - \mu_1}{\sigma}\right) = P(Z \le 0.0171) \approx 0.507.$$

(d) The quantities affected by the change from $\sigma = 5.85$ to 10.0 are p_2, q_1 and p_3. The new values are

$$p_2' = P(X_3 \le 172) = P\left(\frac{X_3 - \mu_2}{\sigma_{new}} \le \frac{172 - \mu_2}{\sigma_{new}}\right) = P(Z \le -0.71) \approx 0.239,$$

$$q_1' = X_{p_1} = \mu_2 + Z_{p_1}\sigma_{new} \approx 179.1 + (-0.5993) \times 10.0 = 173.11.$$

$$p_3' = P(X_1 \le q_1') = P\left(\frac{X_1 - \mu_1}{\sigma} \le \frac{173.11 - \mu_1}{\sigma}\right) = P(Z \le -0.4092) \approx 0.341.$$

The following R script can be used to calculate the answers:

```
> ## (a)
>
> p1 = pnorm(172, mean=175.5, sd=5.84)
> p2 = pnorm(172, mean=179.1, sd=5.84)
> p1
[1] 0.2744814
> p2
[1] 0.1120394
>
> ## (b)
>
> q1 = qnorm(p1, mean=179.1, sd=5.84)
> q1
[1] 175.6
>
> ## (c)
```

```
>
> p3 = pnorm(q1, mean=175.5, sd=5.84)
> p3
[1] 0.5068309
>
> ####### (d)
>
> ## (a)
>
> pp1 = pnorm(172, mean=175.5, sd=5.84)
> pp2 = pnorm(172, mean=179.1, sd=10.0)
> pp1
[1] 0.2744814
> pp2
[1] 0.2388521
>
> ## (b)
>
> qq1 = qnorm(pp1, mean=179.1, sd=10.0)
> qq1
[1] 173.1068
>
> ## (c)
>
> pp3 = pnorm(qq1, mean=175.5, sd=5.84)
> pp3
[1] 0.3409814
>
```

9. In *Natural Inheritance* by Francis Galton, published in 1889, the paired heights of parents with their adult children were reported. The heights, in inches, of 928 children are summarized in the following table. Essentially, we have a histogram. For example, 165 of the 928 adult children have heights in the class interval $(64.7, 66.7]$.

   We are interested in determining whether or not a normal distribution would be appropriate for modeling these heights. We can extract from the table estimates of mean and standard deviation $\mu \approx 68.1$ and $\sigma \approx 2.60$ (by assuming that each datum is represented by the midpoint of its class interval). In principle, we could use the empirical rule (Section 10.6) to assess the normality of the data, except for the fact that we could not expect the quantities $\mu \pm K\sigma$, $K = 1, 2, 3$, to land exactly on the endpoints, which we would need in order to obtain the relevant empirical frequencies.

   Of course, we can use other quantities to achieve the same goal. For example, we can obtain directly from the table estimates of the CDF $P(X \leq x)$ for each endpoint $x = 60.7, 62.7, \ldots, 74.7$

and compare them directly to the values predicted by the normal distribution, that is, $P(Y \leq x)$ where $Y \sim N(\mu = 68.1, \sigma^2 = 2.60^2)$. Try this, by filling in the table. See Section 4.4.4 of *Biostatistics: A Methodology for the Health Sciences* [L.D. Fisher & G. van Belle] for more on this problem.

Class Interval	Freq.	Cumulative Freq.	Estimated CDF	CDF Predicted by Normal Distribution
(58.7, 60.7]	0	-	-	-
(60.7, 62.7]	12	-	-	-
(62.7, 64.7]	91	-	-	-
(64.7, 66.7]	165	-	-	-
(66.7, 68.7]	258	-	-	-
(68.7, 70.7]	266	-	-	-
(70.7, 72.7]	105	-	-	-
(72.7, 74.7]	31	-	-	-

SOLUTION

The requires numbers can be calculated with the following `R` script:

```
> x0 = c(58.7, 60.7,62.7,64.7,66.7,68.7,70.7,72.7)
> x = c(60.7,62.7,64.7,66.7,68.7,70.7,72.7,74.7)
> c1 = paste('(',x0,',',x,']', sep='')
> y = c(0,12,91,165,258,266,105,31)
> tab = data.frame(c1,y,cumsum(y),round(cumsum(y)/sum(y),3),
    round(pnorm(x, mean=68.1, sd=2.6),3))
> names(tab) = paste('column',1:5)
> tab
     column 1 column 2 column 3 column 4 column 5
1 (58.7,60.7]        0        0    0.000    0.002
2 (60.7,62.7]       12       12    0.013    0.019
3 (62.7,64.7]       91      103    0.111    0.095
4 (64.7,66.7]      165      268    0.289    0.295
5 (66.7,68.7]      258      526    0.567    0.591
6 (68.7,70.7]      266      792    0.853    0.841
7 (70.7,72.7]      105      897    0.967    0.962
8 (72.7,74.7]       31      928    1.000    0.994
>
```

This gives:

Class		Cumulative	Estimated	CDF Predicted by
Interval	Freq.	Freq.	CDF	Normal Distribution
(58.7,60.7]	0	0	0.00	0.00
(60.7,62.7]	12	12	0.01	0.02
(62.7,64.7]	91	103	0.11	0.10
(64.7,66.7]	165	268	0.29	0.29
(66.7,68.7]	258	526	0.57	0.59
(68.7,70.7]	266	792	0.85	0.84
(70.7,72.7]	105	897	0.97	0.96
(72.7,74.7]	31	928	1.00	0.99

The estimated cumulative frequencies are reasonably close to the frequencies predicted by the normal distribution.

10. The file `statepop.csv` (posted on BLACKBOARD) is a comma delimited text file with 50 records, each consisting of a state name and the state's population (June 2014).

    (a) Read the file into the `R` environment using the `read.table()` function.

    (b) For each state, calculate the population proportion.

    (c) Rank the proportions in decreasing order, then construct a *log-log* plot, as in Figures 4.11-4.12 of the lecture notes.

    (d) Using the `lines()` function, superimpose on this plot the lines

    $$f(k) = \log(p_X(1)) - \alpha \log(k),$$

    for $\alpha = 1/3, 2/3, 1$, where k represents the rank of a frequency ($k = 1$ is the largest frequency), and $-\alpha$ is an exponent of a power law. Use the `legend()` function to label the superimposed lines. To do this, select distinct `lty` parameters for the 3 superimposed lines, then use these values in the `legend()` function.

    (e) Do state population sizes conform to a power law? Note that the power law may hold for the largest frequencies, but not the smallest.

SOLUTION

The following `R` script produces Figure 1 below:

```
> tab = read.table(’statepop.csv’,header=FALSE,stringsAsFactors=FALSE,sep=’,’)
> names(tab) = c("state","pop")
>
> freq = tab$pop/sum(tab$pop)
>
> y = log(rev(sort(freq)))
>
> par(mfrow=c(1,1),cex=1)
> n = 50
> plot(log(1:n),y,xlab = 'log rank', ylab = 'log frequency')
> alpha = 1
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=1)
> alpha = 2/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=2)
> alpha = 1/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=3)
> legend('bottomleft',legend=paste('alpha = ',c('1','2/3','1/3'),sep=''),lty=1:3)
```

The highest ranking frequencies appear to conform to a power law $p_X(k) \propto 1/k^{2/3}$.

11. Suppose a casino has a game in which a player bets x dollars, then with probability p wins back $2x$ dollars (for a net gain of x) and loses the original x dollars with probability $1-p$ (for a net loss of x). Usually, $p < 1/2$. If $p = 1/2$ then the game is *fair*. Probability theory states that in such a fair game, there can be no strategy that results in a positive expected gain.

    A commonly claimed counter-example to this is the following strategy. Enter the casino, then play the game, betting $x = 1$ each time, until you have a total gain of 1. For example, the following Win/Loss sequence will accomplish this: *LWLLWWW*, which has gain sequence -1,0,-1,-2,-1,0,1, taking 7 games to reach a gain of 1. The Win/Loss sequence *W* also achieves a gain of 1 after a single game. Probability theory also states that the probability that a gain of 1 is reached after a finite number of games is 1 (although this *doesn't* hold if $p < 1/2$).

    This seems to lead to a contradiction, since if we use this strategy, we can play once a day, and guarantee ourselves a regular income, noting that we can use any value of x we wish. Note that the case of the fair game, $p = 1/2$, is the important one, since if no winning strategy exists for this case, no winning strategy can exist when $p < 1/2$, which settles the matter.

    (a) Write an `R` program which simulates this process. Assume $p = 1/2$. For a single simulation, start at *gain* $= 0$, then increase or decrease *gain* after each game by 1. This is a *random walk*. The process stops when *gain* $= 1$. Store the number of games T needed to reach *gain* $= 1$. You may use the `rbinom()` function. Truncate the process at 1000 games. If *gain* $= 1$ has not been reached, indicate this by setting the number of games at, say, $T = 1001$.

    (b) Repeat the simulation to get 1000 replicates of T. Estimate the PMF $p_T(k) = P(T = k)$ directly from the data. Construct a *log-log* plot of $\log(p_T(k))$ against $\log(k)$. Note that the frequencies are not sorted in this case. How many times did T exceed 1000? How many times was T within 10 games, inclusive?

    (c) Using the `lines()` function, superimpose on this plot the lines

    $$f(k) = \log(p_X(1)) - \alpha \log(k),$$

    for $\alpha = 1.0, 1.25, 2.0$. Label your plot with the `legend()` function as in Question 4. If you can conclude that T conforms to a power law, what can be said about $E[T]$? The rate at which this strategy earns money is

$$\text{gain rate} = \frac{\textit{gain}}{\text{number of games played}}.$$

At what rate does this strategy earn money?

SOLUTION

The following R script produces the plot in Figure 2.

```
> nsim = 1000
> tt = rep(NA, nsim)
> bank = rep(NA, nsim)
>
> for (i in 1:nsim) {
+
+  x = rbinom(1000,size=1,prob=1/2)
+  z = cumsum(2*x-1)
+
+  if (sum(z==1) > 0) {
+  tt[i] = min(which(z==1))
+  bank[i] = min(z[1:tt[i]])
+  }
+  else
+  {
+  tt[i] = 1001
+  bank[i] = min(z)
+  }
+ }
>
> sum(tt <= 10)
[1] 752
> sum(tt == 1001)
[1] 32
>
> xx = as.integer(names(table(tt)))
> par(mfrow=c(1,1),cex=1)
> plot(log(xx), log(table(tt)/1000),xlab = ’log T’, ylab = ’log frequency’)
> lines(log(xx), max(log(table(tt)/1000)) - 1*log(xx),lty=1)
> lines(log(xx), max(log(table(tt)/1000)) - 1.25*log(xx),lty=2)
> lines(log(xx), max(log(table(tt)/1000)) - 2*log(xx),lty=3)
> legend(’bottomleft’,legend=paste(’alpha = ’,c(’1.0’,’1.25’,’2.0’),sep=’’),lty=1:3)
>
```

In this simulation there were 752/1000 simulated values of T within 10, and 32/1000 greater than 1000. Results will vary. From Figure 2 the power law $p_X(k) \propto 1/k^\alpha$ holds approximately, with $\alpha \approx 1.25$, and more generally with $\alpha < 2$. We then have, for some constant c,

$$E[T] = \sum_{k=1}^{\infty} k \frac{c}{k^\alpha}$$

noting that the support of T is unbounded. However, $E[T] < \infty$ only if $\alpha > 2$ (compare the summation to the integral $\int_1^\infty x^{-\alpha} dx$). If $\alpha < 2$ then in our case $E[T] = \infty$. This means that although the gambler can win a gain of 1 with probability 1 in a finite amount of time, the *gain rate* is 0, since $E[T] = \infty$. If we let G_n be the total gain after the nth game (not day), we would find

$$\lim_{n \to \infty} \frac{G_n}{n} = 0.$$

As a practical matter, using this strategy, we would often find that we cannot play enough games in a single day to achieve the daily gain of 1.

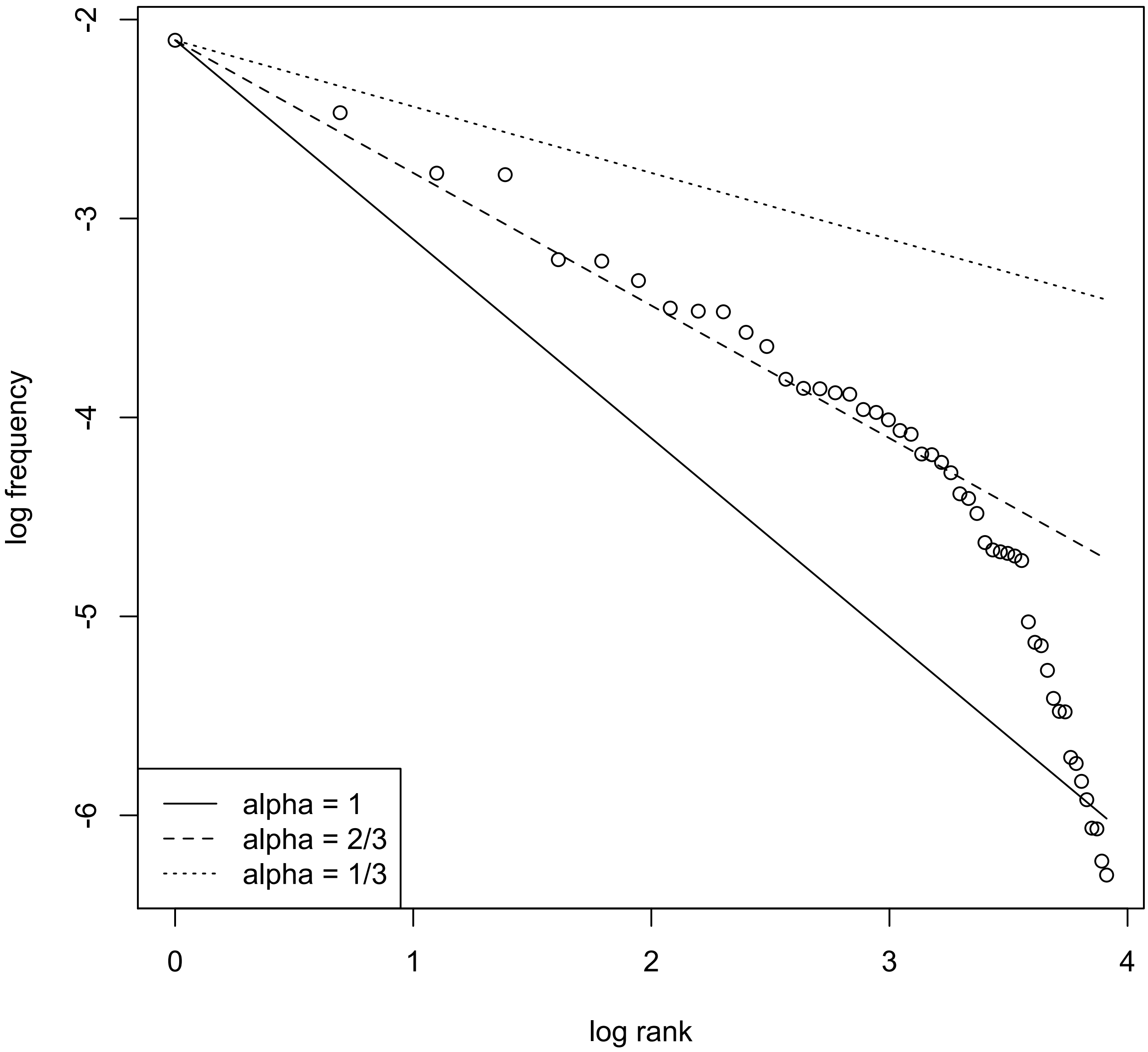

Figure 1: Plot for Problem 6.

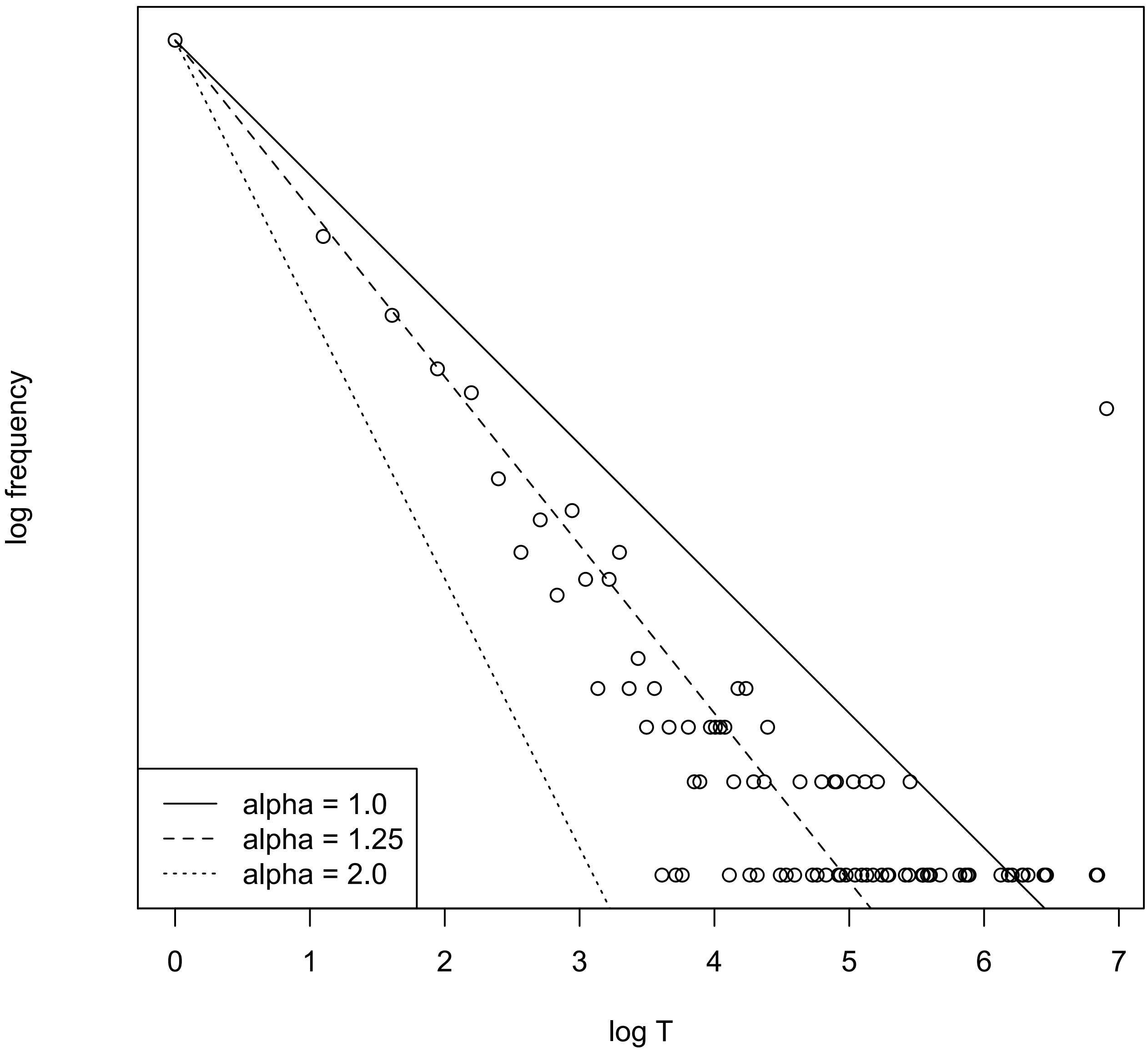

Figure 2: Plot for Problem 7.